BigVGAN: A Universal Neural Vocoder with Large-Scale Training

Abstract

Despite recent progress in generative adversarial network (GAN)-based vocoders, where the model generates raw waveform conditioned on acoustic features, it is challenging to synthesize high-fidelity audio for numerous speakers across various recording environments. In this work, we present BigVGAN, a universal vocoder that generalizes well for various out-of-distribution scenarios without fine-tuning. We introduce periodic activation function and anti-aliased representation into the GAN generator, which brings the desired inductive bias for audio synthesis and significantly improves audio quality. In addition, we train our GAN vocoder at the largest scale up to 112M parameters, which is unprecedented in the literature. We identify and address the failure modes in large-scale GAN training for audio, while maintaining high-fidelity output without over-regularization. Our BigVGAN, trained only on clean speech (LibriTTS), achieves the state-of-the-art performance for various zero-shot (out-of-distribution) conditions, including unseen speakers, languages, recording environments, singing voices, music, and instrumental audio. We release our code and model at: https://github.com/NVIDIA/BigVGAN